Bias and Variance in Multiparty Election Polls

Peter Selb, Sina Chen, John Körtner, and Philipp Bosch

May 16, 2023

Abstract

Recent polling failures highlight that election polls are prone to biases that the margin of error customarily reported with polls does not capture. However, such systematic errors are difficult to assess against the background noise of sampling variance. Shirani-Mehr et al. (2018) developed a hierarchical Bayesian model to disentangle random and systematic errors in poll estimates of two-party vote shares at the election level. The method can inform realistic assessments of poll accuracy. We adapt the model to multiparty elections and improve its temporal flexibility. We then estimate bias and variance in 5,240 German national election polls 1994–2021. Our analysis suggests that the average absolute election-day bias per party was about 1.5 p.p., ranging from 0.9 for the Greens to 3.2 for the Christian Democrats. The estimated variance is, on average, about twice as large as that implied by usual margins of error. We find little evidence of house or mode effects. Common biases indicate industry effects due to similar polling methods. The Supplementary Material provides additional results for 1,751 regional election polls.

1 Introduction

Investigations into alleged polling misses such as the 2020 and 2016 US presidential races (Clinton et al., 2021; Kennedy et al., 2018) or the 2015 UK general election (Sturgis et al., 2018) document the many error sources that can distort polls: unrepresentative samples, failures to adjust for coverage and nonresponse problems, misspecified likely voter models, and misreporting of (or late swings in) vote intentions. The *margin of error* that is regularly reported to convey a poll's uncertainty does not reflect any of these errors – it solely captures the random fluctuation of an estimator across hypothetical replications of the sampling process. Statistical theory and design information suffice to estimate such variance from a single probability sample. Assessing non-sampling errors is far less convenient as it requires validated records, population benchmarks, and possibly empirical replications of the sampling and measurement process. The proliferation of election polling over the past decades has created perhaps the only opportunity in the realm of survey research which brings us close to *observing* both the sampling distribution of estimates from replications of approximately the same sampling process, as well as the underlying population parameter (i.e., the election result).

Most meta studies of polling errors only look at the *total survey error*, i.e., the overall discrepancy between a poll and the election result, and thus confound sampling and non-sampling errors (e.g., Crespi, 1988; Jennings and Wlezien, 2018). An early exception that tries to decompose total errors is Buchanan (1986), who analyzes 155 polls covering 68 elections from nine countries. Due to the small numbers of polls per election (2.3 on average), Buchanan limits his analysis to the top two competitors in each election and pools all 155 two-party estimates assuming (quite daringly!) that they came from the same sampling distribution. Taking expectations, Buchanan finds poll bias to the detriment of conservative parties and variance more than twice as large as what sampling theory implies for simple random samples (SRS). Schnell and Noack (2014) in their study of 145 German Bundestag election polls, 1957-2013, refine the margin of error for each poll to account for multiple parties and common design deviations from SRS (i.e., cluster and stratified sampling). The authors interpret coverage probabilities of their adjusted margins well below the nominal 95% level as indications of bias, but they do not directly estimate bias (nor variance, for that matter). Rather than completely pooling the polls or looking at each poll separately, Shirani-Mehr et al. (2018) take advantage of the multilevel data structure of polls nested in elections. They develop a Bayesian statistical model to estimate election-level variance and bias in polled two-party support, "borrowing strength" across all polls in all elections. In their empirical analysis of 4,221 polls from 608 US state-level elections between 1998 and 2014, the authors find an average election-level bias of about two percentage points and average election-level variance clearly in excess of that implied by SRS and standard margins of error.

Analyses like these are important for making realistic judgments on the accuracy of election polls and similar surveys. To extend the method's scope beyond the two-party context of US elections, however, it needs to accommodate multiple parties. In the following section, we present the model by Shirani-Mehr and colleagues in more detail before adapting it to multiparty elections. We then describe our data, which includes 5,240 polls from eight elections to the German national parliament (Bundestag), 1994-2021. Next, we present empirical results. The final section discusses the relevance of the approach for election polling and how it is reported to the public. The Supplementary Material (SM) provides additional results for 1,754 polls from 71 regional parliamentary (Landtag) elections, 1994-2021.

2 A model of polling errors in multiparty elections

Shirani-Mehr et al. (2018) model the two-party Republican vote share p_j measured in poll j as a random draw from a normal distribution with mean π_j and variance σ_j^2 (Equation (1)):

$$p_j \sim \text{Normal}(\pi_j, \sigma_j^2), \tag{1}$$

$$\text{logit}(\pi_j) = \text{logit}(P_{r[j]}) + \alpha_{r[j]} + \beta_{r[j]} t_j, \tag{2}$$

$$\sigma_j^2 = \pi_j(1 - \pi_j)/n_j + \phi_{r[j]}^2. \tag{3}$$

In Equation (2) the mean is decomposed into the two-party vote for the Republicans, $P_{r[j]}$, where $r[j]$ identifies the election for poll j, an election specific bias term, $\alpha_{r[j]}$, and a time trend, $\beta_{r[j]} t_j$, to account for accuracy gains in polls as election day approaches. The temporal distance, t_j, is defined as days to election scaled between 0 and 1 (with 0 being election day). The logit scale ensures that estimated vote shares are bound between zero and one. The variance σ_j^2 (Equation (3)) is composed of the analytic sampling variance of a binomial proportion under SRS, $\pi_j(1 - \pi_j)/n_j$, where n_j is the sample size, and *excess variance*, $\phi_{r[j]}^2$, due to clustering, weighting, and other features of the design and analysis of surveys (see Frankel, 2010).

For the multiparty case, we introduce index k to denote parties so that each poll j provides an estimate of each party k's vote share, $p_{k,j}$, with the unit sum constraint, $\sum_{k=1}^{K} p_{k,j} = 1$, where K is the total number of parties. Analogous to Shirani-Mehr et al., we model $p_{k,j}$ as a random draw from a normal distribution parameterized as follows:

$$p_{k,j} \sim \text{Normal}(\pi'_{k,j}, \sigma_{k,j}^2), \tag{4}$$

$$\pi'_{k,j} = \pi_{k,j} / \sum_{k=1}^{K} \pi_{k,j}, \tag{5}$$

$$\log(\pi_{k,j}) = \log(P_{k,r[k,j]}) + \alpha_{1k,r[k,j]} + \alpha_{2k,l[k,j]} + \sum_{m=1}^{M} \left(\beta_{k,r,m[k,j,m]} B_m(t_j)\right), \tag{6}$$

$$\log(\sigma_{k,j}^2) = \log(\pi'_{k,j}(1 - \pi'_{k,j})/n_j) + \phi_{k,r[k,j]}. \tag{7}$$

There are some differences to Shirani-Mehr et al.'s (2018) approach. First, we use the inverse multinomial link function in Equation (5) to ensure that the K estimated party vote shares for each poll j are positive and sum to unity. [1] Second, we model the log of $\pi_{k,j}$ in Equation (6) as a function of the log of the party's actual vote share in election r, $P_{k,r[k,j]}$, and two bias terms to capture both party-specific discrepancies between polls and elections, $\alpha_{1k,r[k,j]}$, and *house effects*, $\alpha_{2k,l[k,j]}$, due to differences in survey methods between polling institutes or herding (e.g., Jackman, 2005). [2] To identify the parameters in Equation (6), we use a redundant parametrization, hence estimating $\alpha_{1k,r[k,j]}$, $\alpha_{2k,l[k,j]}$ and $\beta_{k,r,m[k,j,m]}$ only for $K-1$ parties while fixing the parameter for the last category at 0. Third, we include all the polls conducted during the whole legislative periods, whereas Shirani-Mehr et al. include only those in the three weeks before election day. Therefore, possible shifts in electoral mood are a major concern. To that end, we split each legislative period into deciles and specify cubic B-splines (e.g., Green and Silverman, 1993), $B_m(t_j)$, to account for the marked curvilinear patterns observed in the polls (see Figure 1). We let the M spline coefficients $\beta_{k,r,m[k,j,m]}$ vary by party and election. The inclusion of splines dramatically improves the model fit over linear and other global polynomial specifications, and thus also contributes to a realistic assessment of poll variance about the time trends. Finally, we model the party-specific total variance $\sigma^2_{k,j}$ in Equation (7) as the sum of the analytic sampling variance of a multinomial proportion under SRS, $\pi'_{k,j}(1-\pi'_{k,j})/n_j$, and deviation $\phi_{k,r[k,j]}$. In contrast to Shirani-Mehr et al. (2018), we model the variance on the log scale. That way, $\phi_{k,r[k,j]}$ can be positive (e.g., due to cluster sampling and weighting) or negative (e.g., due to stratification and other uses of auxiliary information during the sampling stage), while ensuring that the overall variance $\sigma^2_{k,j}$ is positive. To account for the negative covariances of multinomial proportions,

[1] We considered alternative logratio transformations often used in the analysis of compositional data (see Aitchison, 1982), but found them not practical for identifying biases the way we do, since transformed vote shares also depend on the reference vote share. An advantage of our approach over the increasingly used multinomial Dirichlet model (e.g., Stoetzer et al., 2019), on the other hand, is that we can estimate separate variances for each party.

[2] Note that the index $k, r[k,j]$ identifies party k and election $r[j]$ for party k in poll j and $k, l[k,j]$ identifies party k and polling institute $l[j]$ for party k in poll j. Hence, there is one $\alpha_{1k,r[k,j]}$ for each party and each election, one $\alpha_{2k,l[k,j]}$ for each party and each institute, and there are M $\beta_{k,r,m[k,j,m]}$ for each party and each election. Further, there is one $\phi_{k,r[k,j]}$ for each party and each election.

the bias parameters $\alpha_{1k,r[k,j]}$ and $\alpha_{2k,l[k,j]}$ are given multivariate normal distributions. The spline coefficient $\beta_{k,r,m[k,j,m]}$ and the variance deviation $\phi_{k,r[k,j]}$ are given univariate normal distributions. The model is estimated using Bayesian inference, with hierarchical priors on all parameters. See the SM section D for details.

The fitted model allows us to estimate several key quantities. First, the *average election-day bias* for party k in election r,

$$\hat{b}_{0k,r} = \frac{1}{J_r} \sum_{j \in S_r} (\pi'_{0k,j} - P_{k,r[k,j]}),$$

where S_r is the set of polls in election r, J_r is the number of polls in r, and $\pi'_{0k,j}$ is the estimated vote share of party k in poll j when extrapolating recent poll trends to election day, i.e., when setting the time to election, t_j, to zero in Equation (6). Note that, as a positive average bias for some party has to be compensated by a negative bias for another, $\sum_k \hat{b}_{0kr} = 0$ within election r. We estimate election-day bias both in directional and absolute terms, $|\hat{b}_{0kr}|$. We also estimate *absolute average election-day relbias*, $\frac{|\hat{b}_{0kr}|}{P_r}$, in order to assess bias relative to party size.

Finally, the *average party-institute effect* is defined as

$$\hat{b}_{k,l} = \frac{1}{J_l} \sum_{j \in S_l} (\pi'_{l_{k,j}} - \pi'_{t_{k,j}}),$$

where S_l is the set of all polls conducted by polling institute l, J_l is the number of polls in l, and $\pi'_{l_{k,j}}$ is obtained by omitting $\alpha_{1k,r[k,j]}$ from Equation (6). The estimated vote share at time t_j, $\pi'_{t_{k,j}}$, is obtained by omitting $\alpha_{1k,r[k,j]}$ and $\alpha_{2k,l[k,j]}$ from Equation (6). In other words, party-institute bias measures the average deviation of party k's in institute $l[j]$'s polls from the general poll trend at t_j.

3 Data

Election polls have been conducted in the Federal Republic of Germany since the inaugural Bundestag election in 1949 (Groß, 2010). However, the number of polls (and institutes) was limited during the first decades, and the party system was subject to major changes up to and including the first election after German re-unification in 1990 (Zittel, 2018). Our cross-election perspective requires some continuity, therefore we limit the main empirical analysis to the eight most recent Bundestag elections, 1994-2021. We consider five major parties: the Christian Democrats (CDU/CSU), the Social Democrats (SPD), the Liberal Democrats (FDP), the Greens (B90/GRUENE), and the Left (Die LINKE). Other parties are lumped into one 'others' category.[3] That is, $K = 6$ in our case. The SM contains separate analyses of the 2013, 2017, and 2021 Bundestag elections in which another major party, the populist Alternative für Deutschland (AfD), emerged. The SM also contains additional results for 1,751 polls on 71 regional (Landtag) elections from 1994 to 2021.

We scraped the polling data from `wahlrecht.de` and `dawum.de`, two independent websites on elections, electoral rules, and voting rights in Germany, which provide a real-time collection of published vote intention surveys from nine prominent polling firms (IfD Allensbach, TNS Emnid/Kantar, Forsa, Forschungsgruppe Wahlen, GMS, Infratest dimap, INSA, YouGov, Civey) since the 1998-2002 legislative period. Earlier polling data were kindly provided by Jochen Groß (2010) through Simon Munzert and his colleagues (Stoetzer et al., 2019). A total of 5,240 polls is included in the analysis, ranging from 135 for the 1994 election to 1,019 in 2021, with an average of 655 polls per election. 11 polls without publication date were excluded. Polls were excluded if they did not collect vote intentions for all five major parties (40 polls). If information on the sample size was missing, the institute's average sample size was imputed (251 polls). Figure 1 gives an overview of the polls included in the analysis and how they varied over the legislative periods.

[3] There were 1 to 4 "other" parties (including the AfD) in the 1994-2021 elections, with an average of 1.5. They garnered between 3 and 19% of the votes (10% on average).

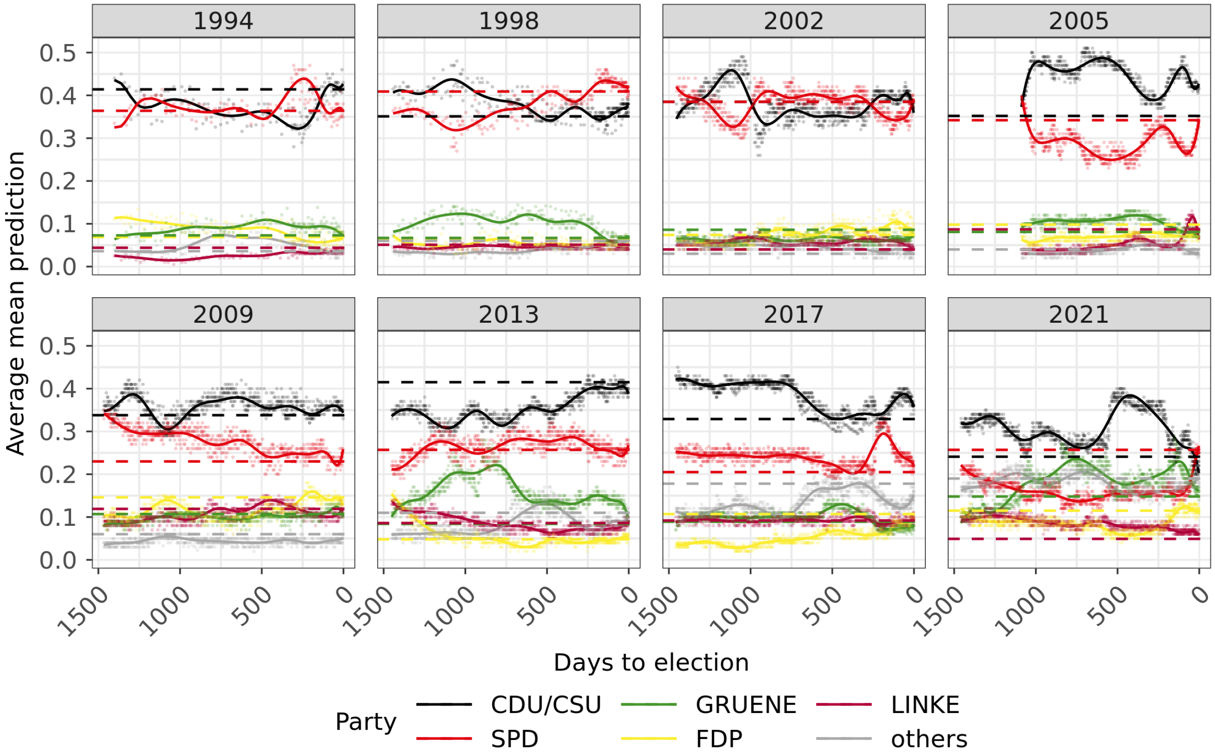

Figure 1: 5,240 national polls by party, Bundestag elections 1994-2021. Dashed horizontal lines indicate election results by party. Curves represent mean posterior predictions (π') of poll shares from our model in Equations (4)-(7). Party-institute effects (α_2) are excluded for clarity of exposition. Note that the 2005 election was called one year early after then-chancellor Gerhard Schröder had lost a motion of confidence in parliament.

Data and code are available from our GitHub repository (`https://github.com/sina-chen/predictors_of_polling_errors`).

4 Empirical results

Figure 2 plots estimated average election-day biases by party and election, 1994-2021. The greatest polling miss occurred in the 2005 election in which all the institutes massively overestimated the CDU/CSU, but nevertheless correctly predicted their victory. Generally, the CDU/CSU appears to be the party which is most difficult to poll. Their average absolute election-day bias is 3.2 p.p. (see Table 1), followed by the FDP and SPD (1.3 p.p.), GRUENE and LINKE (0.9 p.p.), and 'other' parties (1.0 p.p.). These differences lessen if we consider average absolute bias relative to party size (relbias), and there are hardly any other obvious patterns across parties and elections.

Figure 3 presents average party-institute effects for nine polling firms. While tendencies of some institutes to over- or underestimate certain parties are visible here, the effects are generally weak compared to the election-day biases reported above. The patterns that do occur defy common notions of the closeness of some of the institutes to political parties, neither are there any clear differences between survey modes (Prosser and Mellon, 2018). The seemingly weaker house effects among the 'new' institutes which exclusively conduct online surveys (INSA, YouGov, and Civey) may also be an artifact of model-induced smoothing due to their smaller total numbers of polls.

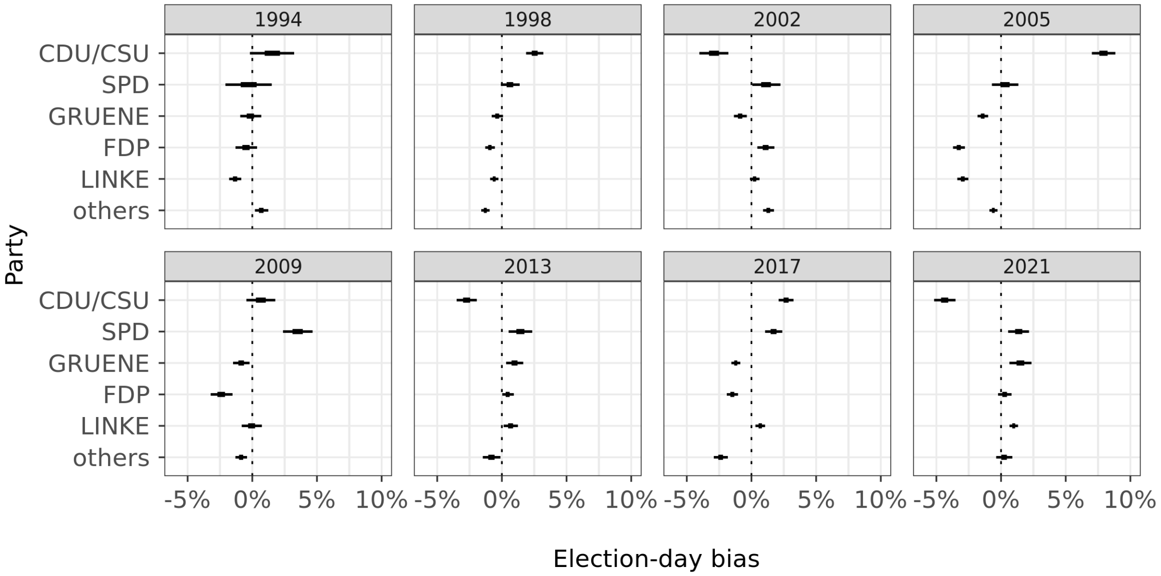

Figure 2: Estimated average election-day biases, $\hat{b}_{0k,r}$. Positive values indicate that polls, on average, would have overestimated a party's vote share on election day and vice versa. Horizontal lines represent 95% and 50% credible intervals.

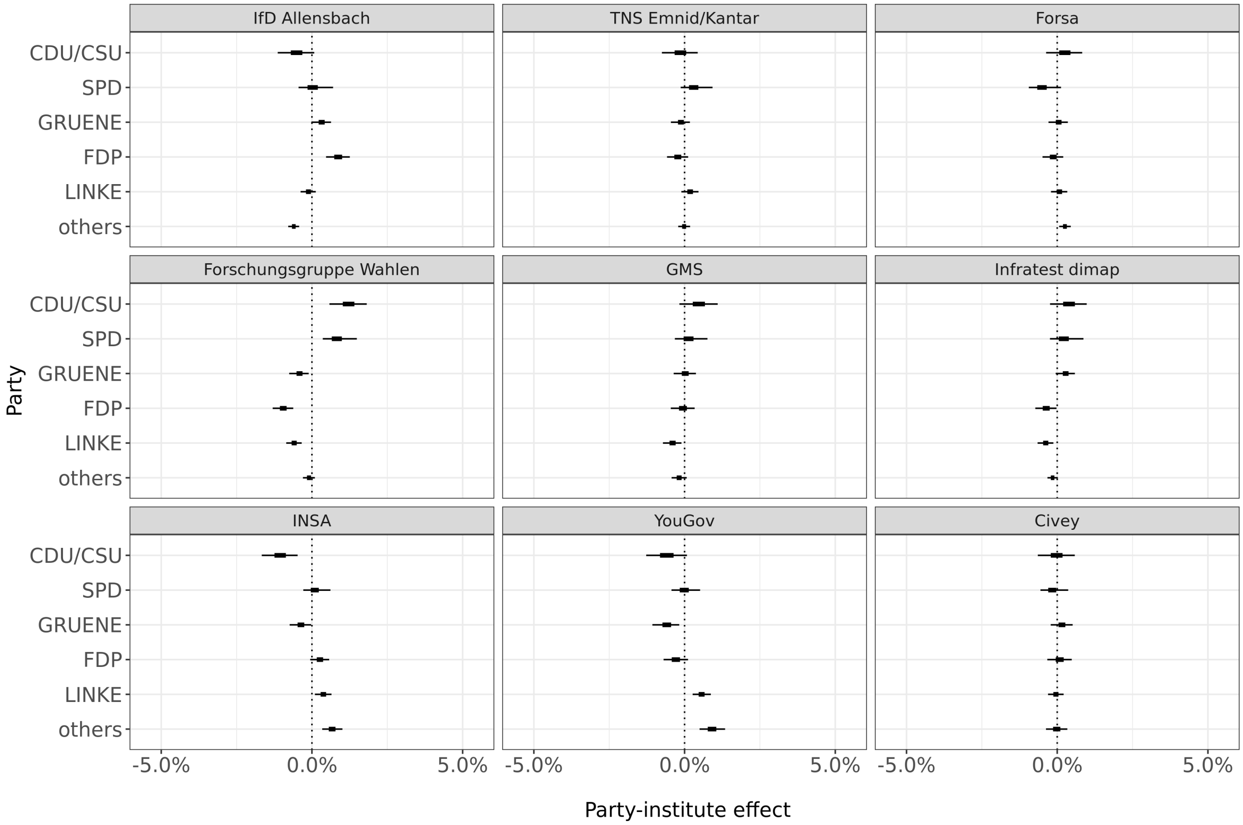

Figure 3: Estimated average party-institute effects, $\hat{b}_{l_k,l}$. Positive values indicate that an institute overestimated a party's vote share relative to the poll average at that time. Horizontal lines represent 95% and 50% credible intervals.

Figure 4 presents estimates of average total standard errors relative to analytic standard errors assuming SRS. In all cases, the total standard error is greater than the analytic standard error, indicating that practical sampling entails efficiency losses compared to SRS. Schnell and Noack (2014) report *design effects*, i.e., the ratio of the sampling variance of estimators based on a given design and an equally sized SRS, for vote intentions in a 2008 German social survey. They consider both geographical sampling points and interviewers as clusters, and find party-specific effects ranging from 1.4 to 2.0. Unfortunately, we are lacking the individual-level data to directly assess design effects, but we ratios of total variances (which possibly include non-sampling errors) to SRS variances in Table 1 for comparison. These are slightly higher, ranging from 1.6 (LINKE) to 3.1 (others), where the latter could also be an artifact of the varying composition of that category. The average variance ratio is at 2.3. Therefore, the SRS margins of error the institutes usually report with published polls underestimate the uncertainty of poll results.

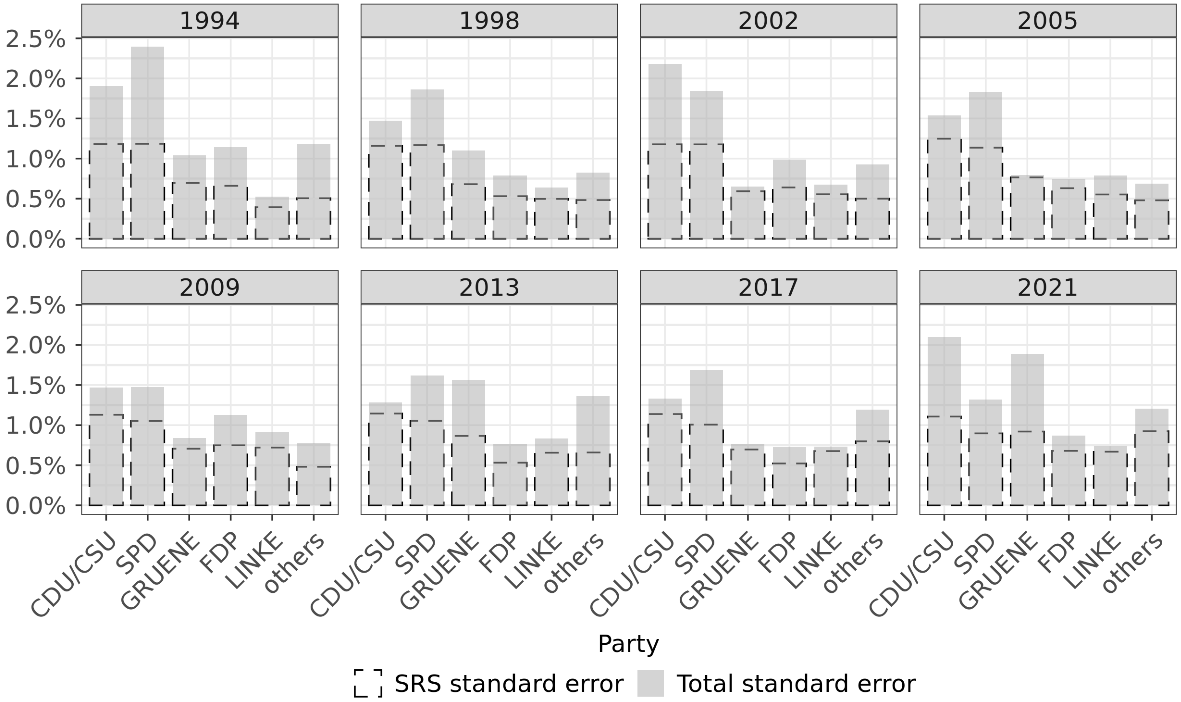

Figure 4: Estimates of average election-level total standard deviation relative to SRS standard deviation.

	CDU/CSU	SPD	GRUENE	FDP	LINKE	others
Absolute election day bias	3.167 (0.491)	1.296 (0.521)	0.918 (0.297)	1.298 (0.292)	0.938 (0.231)	1.020 (0.248)
Absolute election day relbias	0.095 (0.014)	0.051 (0.018)	0.098 (0.033)	0.14 (0.035)	0.147 (0.035)	0.169 (0.039)
SRS variance	0.013 (0.000)	0.012 (0.000)	0.006 (0.000)	0.004 (0.00)	0.004 (0.000)	0.004 (0.000)
Total variance	0.029 (0.002)	0.032 (0.002)	0.013 (0.001)	0.008 (0.001)	0.005 (0.000)	0.011 (0.001)
Variance ratio	2.137 (0.138)	2.627 (0.183)	2.158 (0.140)	2.120 (0.148)	1.570 (0.107)	3.101 (0.223)

Table 1: Mean posterior estimates of absolute election-day bias, relbias, total variance, SRS variance, and variance ratio in Bundestag election polls, 1994-2021. Standard deviation in parentheses.

5 Conclusion

In this paper, we extended the scope of Shirani-Mehr et al.'s (2018) method of disentangling variance and bias in election polls to accommodate multiple parties and volatility in electoral mood. Our empirical analysis of German election polls 1994-2021 largely resonates with what Shirani-Mehr et al. have found for US elections 1998-2014: Average absolute election-day biases between 0.9 and 3.2 p.p., and election-level variances 1.5 to 3 times as large as those implied by standard margins of error. We also looked for house effects but found mostly consistent party-specific biases across polling institutes. Common biases across institutes suggest *industry effects* due to similar polling methods (or herding).

After the 2020 US Presidential election poll miss, academic pollsters have again questioned whether the margin of error is a useful metric, but a suitable alternative is not clear (Schulson, 2020). As Shirani-Mehr et al. (2018) point out, there is little prospect of a general statistical theory of non-sampling errors to inform an alternative metric. In the absence of a convenient analytic measure, however, the convergence of empirical evidence from diverse electoral contexts is reassuring and suggests that there are regularities that can and should be taken into account by pollsters and journalists in realistic assessments of poll accuracy. Information about past mistakes may also be included in model-based election forecasts (Linzer, 2013; Selb and Munzert, 2016). Still, assessing bias and variance in past polls alone does not help to predict the magnitude and direction of errors in current polls. The modelling approach advocated here is easily extended to incorporate predictor variables at various levels to account for contextual correlates of polling misses. Such an extension could thus prospectively identify elections in which substantial polling errors are likely to occur. We recognize that our strategy, which involves a patchwork of distributions and transformations, is somewhat ad hoc in the light of current theory and practice of compositional data analysis (see Pawlowsky-Glahn and Buccianti, 2011, for an overview), but we appreciate its flexibility and simplicity of building it out of linear models.

Besides that application, the ability to distinguish between variance and bias in election polling is crucial to improve election forecasts. If survey errors were in large part random we could try to reduce the variance of estimates, for instance, by increasing sample sizes. This principle is the basic idea underlying *poll aggregators* which average over many polls to forecast election outcomes, effectively increasing sample size and reducing sampling variance compared to estimates based on any individual poll covered (see Jackson, 2018, for an overview). If survey errors are systematic, however, little can be gained from poll averaging other than over-confidence in biased estimates.

References

Aitchison, John. 1982. "The statistical analysis of compositional data." *Journal of the Royal Statistical Society: Series B (Methodological)* 44(2):139–160.

Buchanan, William. 1986. "Election predictions: An empirical assessment." *Public Opinion Quarterly* 50(2):222–227.

Clinton, Josh, Jennifer Agiesta, Megan Brenan, Camille Burge, Marjorie Connelly, Ariel Edwards-Levy, Bernard Fraga, Emily Guskin, D. Sunshine Hillygus, Chris Jackson, Jeff Jones, Scott Keeter, Kabir Khanna, John Lapinski, Lydia Saad, Daron Shaw, Andrew Smith, David Wilson and Christopher Wlezien. 2021. Task Force on 2020 Pre-Election Polling: An Evaluation of the 2020 General Election Polls. Technical report American Association of Public Opinion Research.

Crespi, Irving. 1988. *Pre-election polling: Sources of accuracy and error.* Russell Sage Foundation.

Frankel, Martin. 2010. Sampling theory. In *Handbook of survey research*, ed. James D Wright and Peter V Marsden. Emerald Group Bingley, UK pp. 83–138.

Green, Peter J and Bernard W Silverman. 1993. *Nonparametric regression and generalized linear models: a roughness penalty approach.* Crc Press.

Groß, Jochen. 2010. *Die Prognose von Wahlergebnissen: Ansätze und empirische Leistungsfähigkeit.* Springer.

Jackman, Simon. 2005. "Pooling the polls over an election campaign." *Australian Journal of Political Science* 40(4):499–517.

Jackson, Natalie. 2018. The Rise of Poll Aggregation and Election Forecasting. In *Oxford Handbook of Polling and Polling Methods*, ed. Lonna Rae Atkeson and R. Michael Alvarez. Oxford: Oxford University Press.

Jennings, Will and Christopher Wlezien. 2018. "Election polling errors across time and space." *Nature Human Behaviour* 2(4):276.

Kennedy, Courtney, Mark Blumenthal, Scott Clement, Joshua D Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, Lee Miringoff, Kristen Olson, Douglas Rivers et al. 2018. "An evaluation of the 2016 election polls in the United States." *Public Opinion Quarterly* 82(1):1–33.

Linzer, Drew A. 2013. "Dynamic Bayesian forecasting of presidential elections in the states." *Journal of the American Statistical Association* 108(501):124–134.

Pawlowsky-Glahn, Vera and Antonella Buccianti. 2011. *Compositional data analysis: Theory and applications*. John Wiley & Sons.

Prosser, Christopher and Jonathan Mellon. 2018. "The twilight of the polls? A review of trends in polling accuracy and the causes of polling misses." *Government and Opposition* 53(4):757–790.

Schnell, Rainer and Marcel Noack. 2014. "The accuracy of pre-election polling of German general elections." *methods, data, analyses* 8(1):20.

Schulson, Michael. 2020. "In Fallout Over Polls, 'Margin of Error' Gets New Scrutiny." *Undark* .

Selb, Peter and Simon Munzert. 2016. "Forecasting the 2013 German bundestag election using many polls and historical election results." *German Politics* 25(1):73–83.

Shirani-Mehr, Houshmand, David Rothschild, Sharad Goel and Andrew Gelman. 2018. "Disentangling bias and variance in election polls." *Journal of the American Statistical Association* 113(522):607–614.

Stoetzer, Lukas F, Marcel Neunhoeffer, Thomas Gschwend, Simon Munzert and Sebastian

Sternberg. 2019. “Forecasting elections in multiparty systems: a Bayesian approach combining polls and fundamentals.” *Political Analysis* 27(2):255–262.

Sturgis, Patrick, Jouni Kuha, Nick Baker, Mario Callegaro, Stephen Fisher, Jane Green, Will Jennings, Benjamin E Lauderdale and Patten Smith. 2018. “An assessment of the causes of the errors in the 2015 UK general election opinion polls.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181(3):757–781.

Zittel, Thomas. 2018. Electoral systems in context: Germany. In *The Oxford Handbook of Electoral Systems*, ed. Erik S. Herror, Robert Pekkanen and Matthew S. Shugart. Oxford University Press Oxford pp. 781–801.

Supplementary Material

A 2013 to 2021 Bundestag elections: Detailed results

In this section, we provide detailed results for the 2013 to 2021 Bundestag elections, reporting bias and variance estimates for the Alternative für Deutschland (AfD) separately from the "others" category which we formed in the main text to facilitate cross-election analyses. Founded in February 2013, the AfD narrowly failed to clear the 5 percent threshold in the Bundestag elections in September 2013. After several successful regional (Landtag) elections, the AfD entered the Bundestag in the 2017 election with 12.6 percent of the vote and has since gained representation in all regional parliaments, with vote shares ranging from 5.9% (Schleswig-Holstein 2017) to 27.5% (Saxony 2019). Figure 5 tracks 2,064 polls over the election cycles 2009-2013, 2013-2017, and 2017-2021 included in the analysis, with the "others" category from the main text broken down to distinguish the AfD from minor parties. Note that the shortened observation period for the 2013 Bundestag election is due to the AfD entering the race only half a year before election day. The underestimation of anti-establishment parties and candidates is a frequently observed phenomenon and is often ascribed to socially desirable survey responses and selective participation in polls (e.g., Kennedy et al., 2018). According to our statistical model, this observation is only partially supported: the average election-day bias for the AfD is -0.4% in 2013, -1.9% in 2017 and +0.5% in 2021. Figure 8 shows that the election-level standard errors are estimated to be somewhat greater than what SRS theory would have one expect, except for the CDU/CSU in 2013. A note of caution is due, though. Like other log-ratio models for compositional data, our model requires *independence of irrelevant alternatives.* In our case this would imply that a new party entering the electoral arena (such as the AfD in 2013) evenly reduces the vote shares of existing parties so that their ratios remain unaffected. This assumption is unlikely to hold in multiparty contests in which different parties compete for different groups of voters (e.g., Alvarez and Nagler, 2000). While this will mostly affect regression coefficients, the results with and without AfD should not be expected to be compatible.

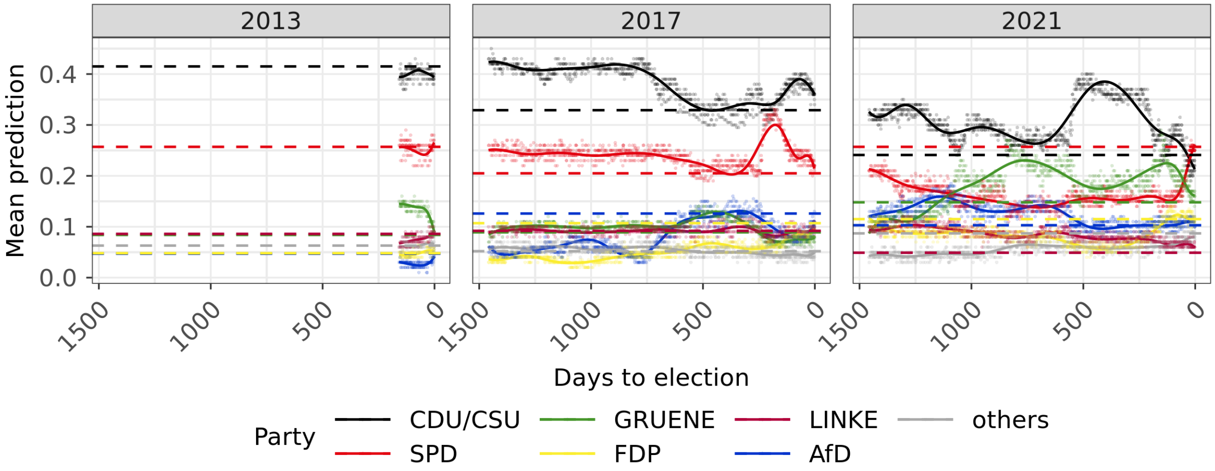

Figure 5: 2,064 national polls by party, Bundestag elections 2013-2021. Dashed horizontal lines indicate actual party vote shares. Curves represent mean posterior predictions (π') of poll shares from our model in (4)-(7). Party-institute effects ($\alpha_{2k,l[k,j]}$) are excluded for clarity of exposition

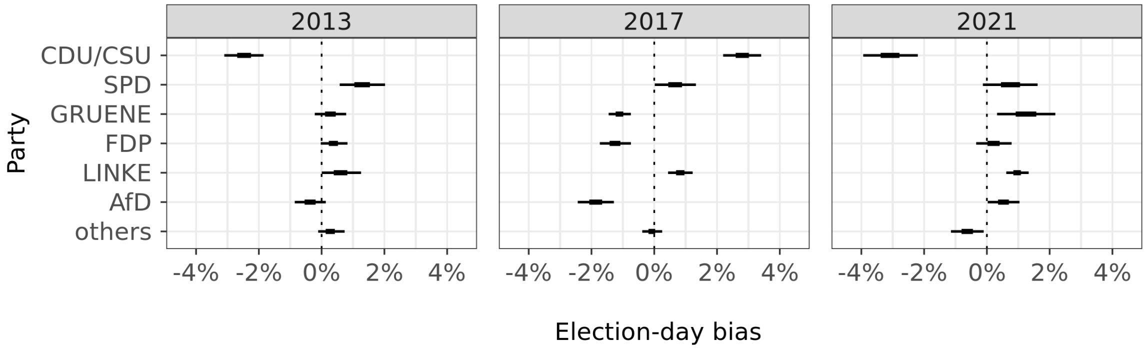

Figure 6: Estimated average election-day biases, $\hat{b}_{0k,r}$. Positive values indicate that polls, on average, would have overestimated a party's vote share on election day and vice versa. Horizontal lines represent 95% and 50% credible intervals.

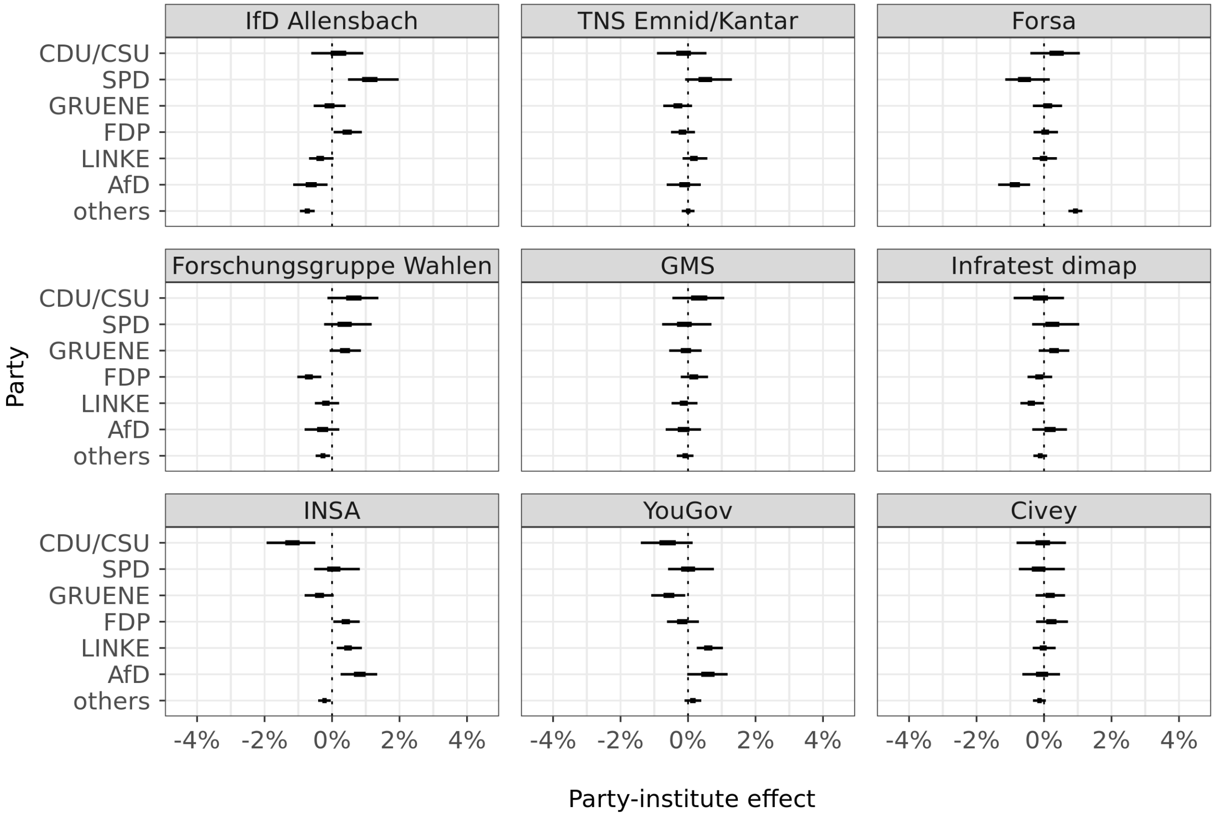

Figure 7: Estimated average party-institute effects, $\hat{b}_{l_{k},l}$. Positive values indicate that an institute overestimated a party's vote share relative to the poll average at that time. Horizontal lines represent 95% and 50% credible intervals.

	CDU/CSU	SPD	GRUENE	FDP	LINKE	AfD	others
Absolute election day bias	2.784 (0.359)	0.904 (0.381)	0.877 (0.303)	0.615 (0.252)	0.800 (0.233)	0.921 (0.271)	0.327 (0.216)
Absolute election day relbias	0.091 (0.012)	0.037 (0.016)	0.081 (0.028)	0.071 (0.031)	0.119 (0.032)	0.092 (0.034)	0.044 (0.032)
SRS variance	0.013 (0.000)	0.001 (0.000)	0.007 (0.000)	0.003 (0.000)	0.004 (0.000)	0.004 (0.000)	0.003 (0.000)
Total variance	0.023 (0.001)	0.018 (0.001)	0.018 (0.001)	0.006 (0.000)	0.006 (0.001)	0.008 (0.000)	0.006 (0.000)
Variance ratio	1.809 (0.102)	1.897 (0.124)	2.356 (0.154)	1.748 (0.133)	1.425 (0.124)	2.099 (0.163)	1.938 (0.150)

Table 2: Mean posterior estimates of absolute election-day bias, total variance, SRS variance, and the average variance ratios for Bundestag election polls, 2013-2021. Standard deviation in parentheses.

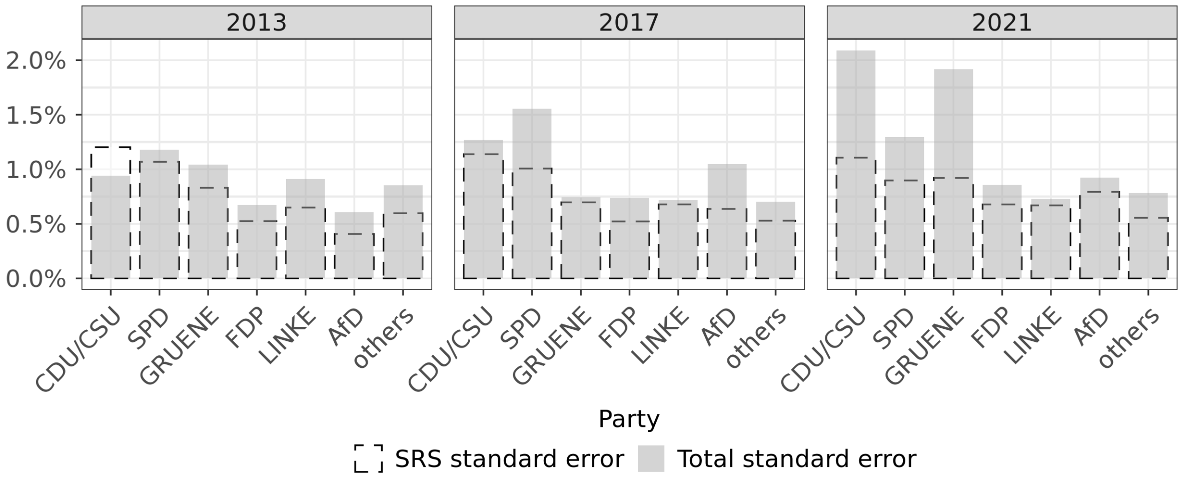

Figure 8: Estimates of average election-level total standard error relative to SRS standard error.

B Regional (Landtag) election polling, 1994-2021.

Germany consists of 16 federal states (Bundesländer) [4], each of which conducts parliamentary (Landtag) elections every four or five years. In this section we provide additional results for 1,751 polls in 71 regional (Landtag) elections from 1994 to 2021. The data were retrieved from `wahlrecht.de` (cutoff date: 2021-11-28). For the sake of comparability across elections, we restrict our analysis to polls which cover the five major parties already considered in the main text. This method led to the exclusion of 15 elections (273 polls) in which one or more out of the five major parties were not competing.[5] As in the main analysis, the institute's mean sample size was imputed (224 polls) if information on the sample size had been missing. If there was no information on the sample size for any of the polls conducted by an institute, a value of $n = 1,000$ was imputed (12 polls). Furthermore, we do not model house effects, since there were over 70 institutes active in polling Landtag elections, with few overlaps between states. Figure 9 gives average election-day bias estimates for each party in each

[4] BW: Baden-Württemberg, BY: Bavaria, BE: Berlin, BB: Brandenburg, HB: Bremen, HH: Hamburg, HE: Hesse, MV: Mecklenburg Western Pomerania, NI: Lower Saxony, NW: Northrhine Westphalia, RP: Rhineland Palatinate, SL: Saarland, ST: Saxony Anhalt, SN: Saxony, SH: Schleswig Holstein, TH: Thuringia.

[5] These were the following elections: BW 1996, BW 2001, BY 1994, BY 1998, BY 2003, HB 1999, HH 2001, HH 2004, HE 1995, HE 1999, HE 2003, NW 2000, SH 1996, SH 2000, RP 2001

election included in the analysis. Overall, absolute election-day bias amounts to an average of 2.7% across parties and elections, ranging from 1.5% for the GRUENE to 3.8% for the CDU/CSU. This is substantively higher as compared to average bias in national polls. Figure 10 presents the average total standard errors relative to SRS standard errors.

	CDU/CSU	SPD	GRUENE	FDP	LINKE	others
Absolute election day bias	3.809 (2.202)	3.123 (2.141)	1.485 (0.929)	2.21 (0.681)	1.775 (0.995)	2.399 (0.740)
Absolute election day relbias	0.129 (0.072)	0.14 (0.082)	0.182 (0.107)	0.354 (0.121)	0.167 (0.103)	0.254 (0.079)
SRS variance	0.022 (0.000)	0.02 (0.000)	0.009 (0.000)	0.006 (0.000)	0.010 (0.000)	0.009 (0.000)
Total variance	0.030 (0.012)	0.031 (0.012)	0.015 (0.005)	0.008 (0.003)	0.015 (0.006)	0.024 (0.008)
Variance ratio	1.387 (0.528)	1.522 (0.566)	1.547 (0.596)	1.520 (0.601)	1.497 (0.589)	2.929 (1.000)

Table 3: Mean posterior estimates of absolute election-day bias, relbias, total variance, SRS variance, and variance ratio in Landtag election polls, from 1994 to 2021. Standard deviation in parentheses.

C Bayesian estimation

The model was implemented using the Stan platform for statistical modelling and computing (Stan Development Team, 2020). The full Stan model as well as all code are available at `https://github.com/sina-chen/predictors_of_polling_errors`. The Bayesian estimation was performed with three parallel chains and 5,000 iterations each. Half of the iterations of each chain were discarded. 7,500 samples were estimated. There were no divergent transitions. In Figure 11 one can see that the potential scale reduction factor $\widehat{R}$ for all estimated parameters is below 1.05 indicating adequate model fit. More detailed information can be found at `https://pollingerrors.shinyapps.io/multiparty_bias_variance_btw94_21/`.

D Prior specification

To account for the structure of our data, we place hierarchical priors on all parameters. We choose weakly informative priors which allow substantive but not excessive poll bias or variance.

The redundant parametrization of $\alpha_{1k,r[k,j]}$ and $\alpha_{2k,l[k,j]}$ ensures that their covariance matrices are identifiable. Since there is a linear interdependence between the K estimated party parameters for each election or each institute, the covariance matrices might be singular, hence there is no unique inverse and the matrices cannot be estimated. The redundant parametrisation resolves the linear interdependence.

Resulting, priors for $\alpha_{1k,r[k,j]}$ and $\alpha_{2k,l[k,j]}$ are only specified for $K-1$ parties. In the following, the subscripts α_1 and α_2 indicate that the respective hyperprior belongs to $\alpha_{1k,r[k,j]}$ or $\alpha_{2k,l[k,j]}$. For computational efficiency, the $(K-1) \times (K-1)$ covariance matrices $\Sigma_{\alpha 1}$ and $\Sigma_{\alpha 2}$ are obtained by combining the $(K-1) \times (K-1)$ correlation matrices $\Omega_{\alpha 1}$ and $\Omega_{\alpha 2}$ with vectors of the $K-1$ standard deviations $\tau_{\alpha 1}$ and $\tau_{\alpha 2}$ (Barnard, McCulloch and Meng, 2000): $\Sigma_{\alpha 1} = \text{diag}(\tau_{\alpha 1}) \times \Omega_{\alpha 1} \times \text{diag}(\tau_{\alpha 1})$ and $\Sigma_{\alpha 2} = \text{diag}(\tau_{\alpha 2}) \times \Omega_{\alpha 2} \times \text{diag}(\tau_{\alpha 2})$, where $diag(\tau_{\alpha 1})$ and $diag(\tau_{\alpha 12})$ represent $(K-1) \times (K-1)$ matrices with diagonal elements $\tau_{\alpha 1}$ and $\tau_{\alpha 2}$. The correlation matrices $\Omega_{\alpha 1}$ and $\Omega_{\alpha 2}$ follows a LKJ distribution with a hyperprior of 2, which was chosen due to convergency constraints: $\Omega_{\alpha 1} \sim LKJCorr(2)$ and $\Omega_{\alpha 2} \sim LKJCorr(2)$. Instead of parameterising the correlation matrices directly, it is more efficient and numerical stable to decompose them into $\Omega_{\alpha 1} = L_{\alpha 1} \times L'_{\alpha 1}$ and $\Omega_{\alpha 2} = L_{\alpha 2} \times L'_{\alpha 2}$, with $L_{\alpha 1}$ and $L_{\alpha 2}$ being the lower triangular matrices, also known as the Cholesky factor (Stan Development Team, 2020). A half normal distribution is assigned to $\tau_{\alpha 1}$ and $\tau_{\alpha 2}$: $\tau_{\alpha 1} \sim \text{Normal}_{+}(0, \sigma^2_{\tau_{\alpha 1}})$ and $\tau_{\alpha 2} \sim \text{Normal}_{+}(0, \sigma^2_{\tau_{\alpha 2}})$, with their variance specified as: $\sigma^2_{\tau_{\alpha 1}} \sim \text{Normal}_{+}(0, 0.2^2)$, and $\sigma^2_{\tau_{\alpha 2}} \sim \text{Normal}_{+}(0, 0.2^2)$.

The redundantly parameterised $\beta_{k,r,m[k,j,m]}$ is univariate normal distributed for $K-1$ parties: $\beta_{k,r,m[k,j,m]} \sim \text{Normal}(\mu_\beta, \sigma^2_\beta)$, with $\mu_\beta \sim \text{Normal}(0, 0.2^2)$ and $\sigma^2_\beta \sim \text{Normal}_{+}(0, 0.2^2)$. Further, the variance parameter $\phi_{k,r[k,j]}$ is univariate normal distributed: $\phi_{k,r[k,j]} \sim \text{Normal}(0, \sigma^2_\phi)$, with $\sigma^2_\phi \sim \text{Normal}_{+}(0, 0.2^2)$.

References

Alvarez, R Michael and Jonathan Nagler. 2000. "A new approach for modelling strategic voting in multiparty elections." *British Journal of Political Science* 30(1):57–75.

Barnard, John, Robert McCulloch and Xiao-Li Meng. 2000. "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage." *Statistica Sinica* pp. 1281–1311.

Kennedy, Courtney, Mark Blumenthal, Scott Clement, Joshua D Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, Lee Miringoff, Kristen Olson, Douglas Rivers et al. 2018. "An evaluation of the 2016 election polls in the United States." *Public Opinion Quarterly* 82(1):1–33.

Stan Development Team. 2020. "Stan Modeling Language Users Guide and Reference Manual.".

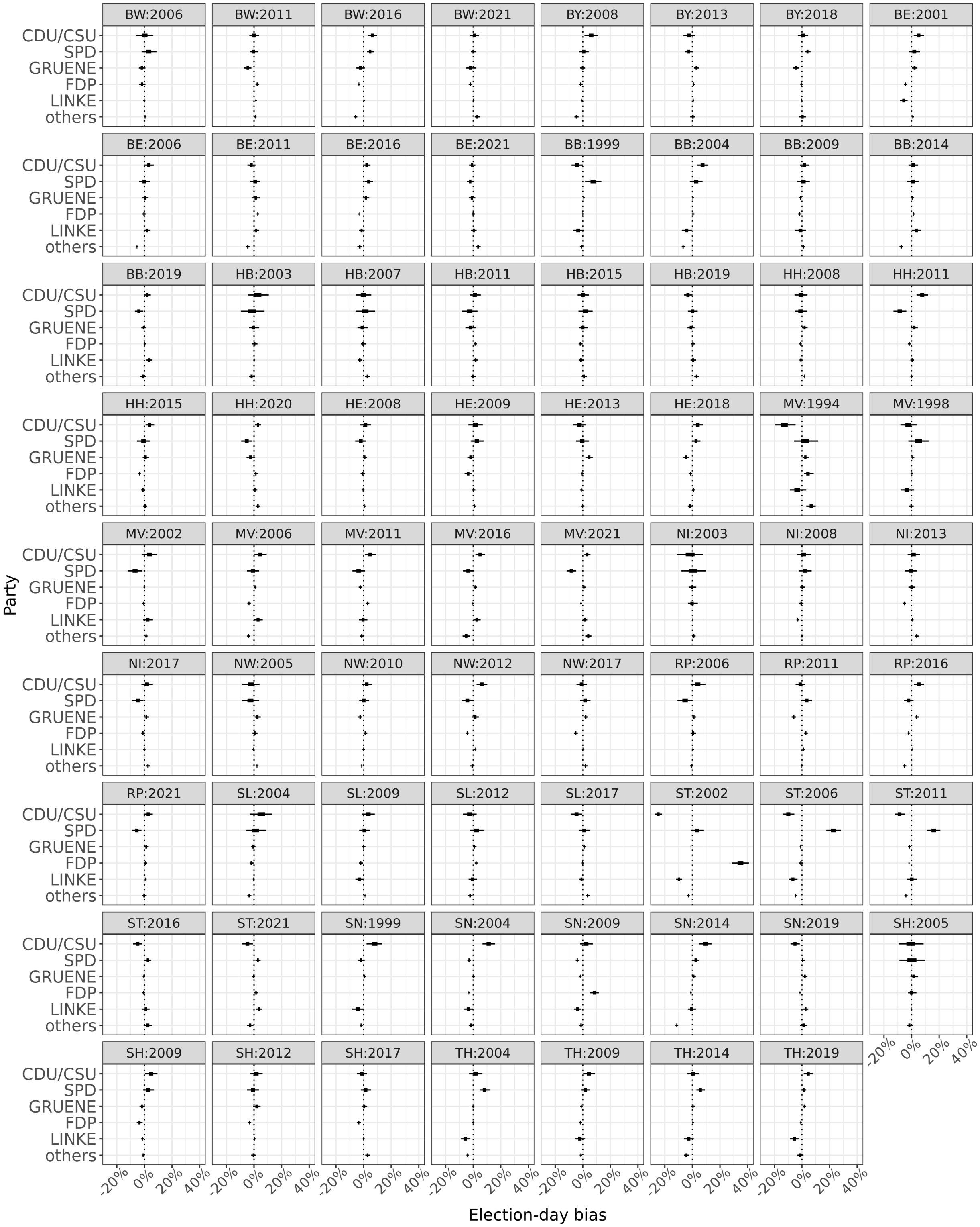

Figure 9: Estimated average election-day biases, $\hat{b}_{0kr}$. Positive values indicate that polls, on average, would have overestimated a party's vote share on election day and vice versa. Horizontal lines represent 95% and 50% credible intervals.

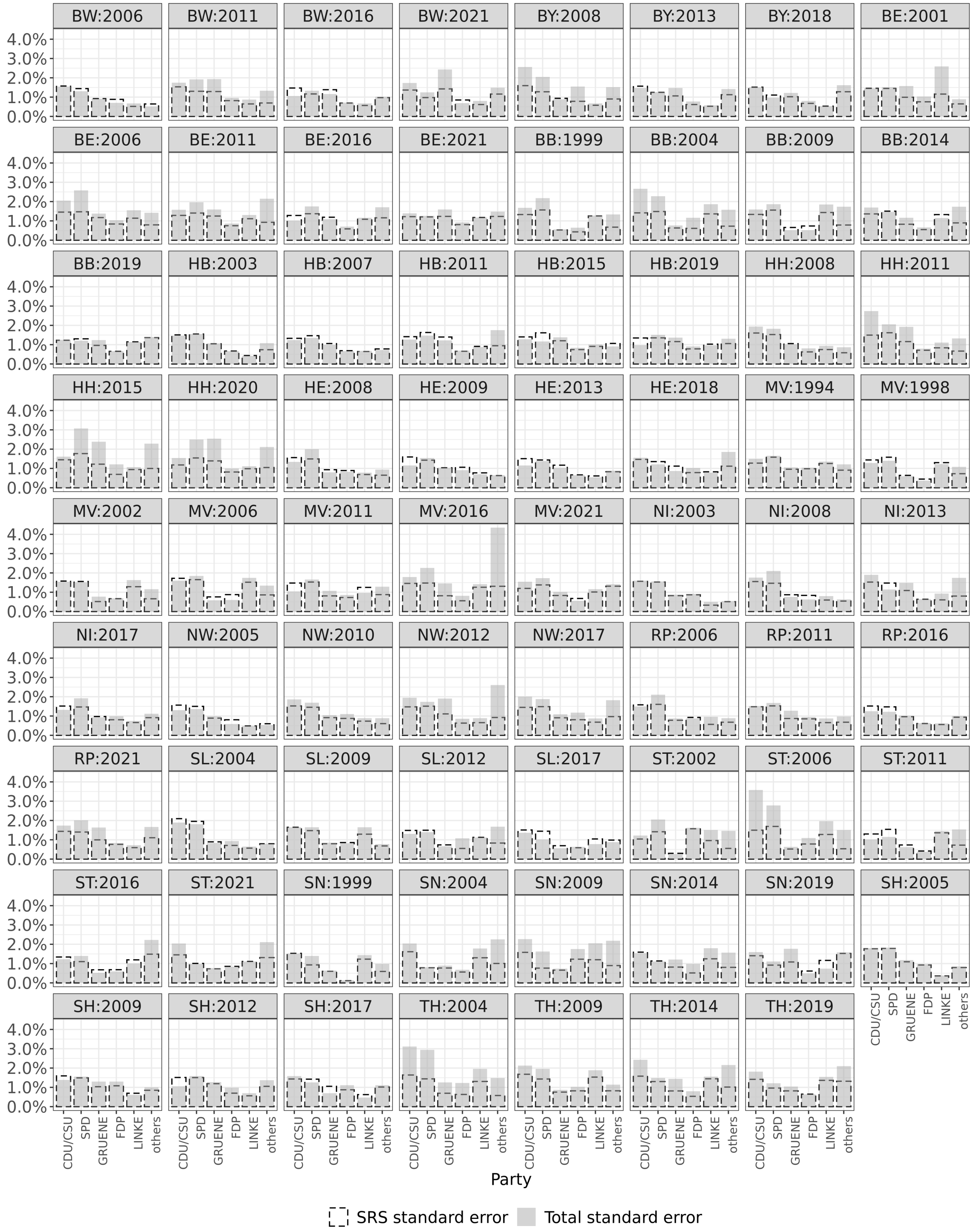

Figure 10: Estimates of average election-level total standard error relative to SRS standard error for German Landtag election polls.

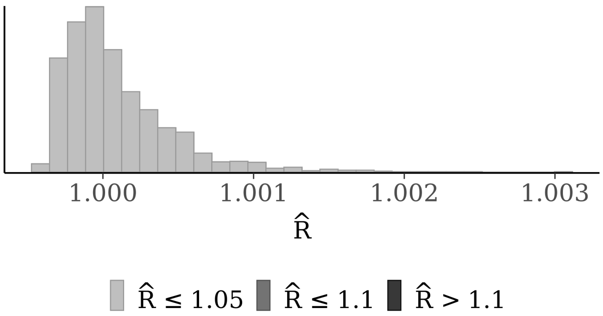

Figure 11: $\widehat{R}$ for German Bundestag election 1994-2021 model.